

# Noisy sorting without resampling

Mark Braverman\*

Elchanan Mossel<sup>†</sup>

February 1, 2008

## Abstract

In this paper we study noisy sorting without re-sampling. In this problem there is an unknown order  $a_{\pi(1)} < \dots < a_{\pi(n)}$  where  $\pi$  is a permutation on  $n$  elements. The input is the status of  $\binom{n}{2}$  queries of the form  $q(a_i, a_j)$ , where  $q(a_i, a_j) = +$  with probability at least  $1/2 + \gamma$  if  $\pi(i) > \pi(j)$  for all pairs  $i \neq j$ , where  $\gamma > 0$  is a constant and  $q(a_i, a_j) = -q(a_j, a_i)$  for all  $i$  and  $j$ . It is assumed that the errors are independent. Given the status of the queries the goal is to find the maximum likelihood order. In other words, the goal is find a permutation  $\sigma$  that minimizes the number of pairs  $\sigma(i) > \sigma(j)$  where  $q(\sigma(i), \sigma(j)) = -$ . The problem so defined is the feedback arc set problem on distributions of inputs, each of which is a tournament obtained as a noisy perturbations of a linear order. Note that when  $\gamma < 1/2$  and  $n$  is large, it is impossible to recover the original order  $\pi$ .

It is known that the weighted feedback arc set problem on tournaments is NP-hard in general. Here we present an algorithm of running time  $n^{O(\gamma^{-4})}$  and sampling complexity  $O_\gamma(n \log n)$  that with high probability solves the noisy sorting without re-sampling problem. We also show that if  $a_{\sigma(1)}, a_{\sigma(2)}, \dots, a_{\sigma(n)}$  is an optimal solution of the problem then it is “close” to the original order. More formally, with high probability it holds that  $\sum_i |\sigma(i) - \pi(i)| = \Theta(n)$  and  $\max_i |\sigma(i) - \pi(i)| = \Theta(\log n)$ .

Our results are of interest in applications to ranking, such as ranking in sports, or ranking of search items based on comparisons by experts.

---

\*C.S. University of Toronto, partially supported by and NSERC CGS scholarship. Part of the work was done while on a visit to IPAM, UCLA

<sup>†</sup>Dept. of Statistics, U.C. Berkeley. Supported by an Alfred Sloan fellowship in Mathematics, by NSF grants DMS-0528488 and DMS-0548249 (CAREER) and by DOD ONR grant N0014-07-1-05-06. Part of this work was done while the author was visiting IPAM, UCLA

# 1 Introduction

We study the problem of sorting in the presence of noise. While sorting linear orders is a classical well studied problem, the introduction of noise poses very interesting challenges. Noise has to be considered when ranking or sorting is applied in many real life scenarios.

A natural example comes from sports. How do we rank a league of soccer teams based on the outcome of the games? It is natural to assume that there is a true underlying order of which team is better and that the games outcome represent noisy versions of the pairwise comparisons between teams. Note that in this problem it is impossible to “re-sample” the order between a pair of teams. As a second example, consider experts comparing various items according to their importance where each pair of elements is compared by one expert. It is natural to assume that the experts opinions represent a noisy view of the actual order of significance. The question is then how to aggregate this information?

## 1.1 The Sorting Model

We will consider the following probabilistic model of instances. There will be  $n$  items denoted  $a_1, \dots, a_n$ . There will be a *true order* given by a permutation  $\pi$  on  $n$  elements such that under the true order  $a_{\pi(1)} < a_{\pi(2)} < \dots < a_{\pi(n-1)} < a_{\pi(n)}$ . The algorithm will have access to  $\binom{n}{2}$  queries defined as follows.

**Definition 1.** For each pair  $i, j$  the outcome of the comparison between  $a_i$  and  $a_j$  is denoted by  $q(a_i, a_j) \in \pm$  where for all  $i \neq j$  it holds that  $q(a_i, a_j) = -q(a_j, a_i)$ . We assume that the probability  $q(a_i, a_j) = +$  is at least  $p := \frac{1}{2} + \gamma$  if  $\pi(i) > \pi(j)$  and that the queries

$$\{q(a_i, a_j) : 1 \leq i < j \leq n\}$$

are independent conditioned on the true order. In other words, for any set

$$S = \{(i(1) < j(1)), \dots, (i(k) < j(k))\},$$

any vector  $s \in \{\pm\}^k$  and  $(i < j) \notin S$  it holds that

$$\mathbb{P}[q(a_i, a_j) = + | \forall 1 \leq \ell \leq k : q(a_{i(\ell)}, a_{j(\ell)}) = s_\ell] = \mathbb{P}[q(a_i, a_j) = +]. \quad (1)$$

It is further assumed that  $1/2 < p = \frac{1}{2} + \gamma < 1$ .

We will be interested in finding a ranking that will minimize the number of upsets. More formally:

**Definition 2.** Given  $\binom{n}{2}$  queries  $q(a_i, a_j)$  the score  $s_q(\sigma)$  of a ranking (permutation)  $\sigma$  is given by

$$s_q(\sigma) = \sum_{i,j:\sigma(i)>\sigma(j)} q(a_{\sigma(i)}, a_{\sigma(j)}). \quad (2)$$

We say that a ranking  $\tau$  is optimal for  $q$  if  $\tau$  is a maximizer (2) among all ranking.

The Noisy Sorting Without Resampling (NSWR) problem is the problem of finding an optimal  $\tau$  given  $q$  assuming that  $q$  is generated as in Definition 1.

The problem of maximizing (2) without any assumptions on the input distribution is called the *feedback arc set problem for tournaments* which is known to be NP-hard, see subsection 1.2 for references, more background and related models.

The score (2) has a clear statistical interpretation in the case where each query is answered correctly with probability  $p$  exactly. In this case, for each permutation  $\sigma$  we can calculate the probability  $P[q|\sigma]$  of observing  $q$  given that  $\sigma$  is the true order. It is immediate to verify that  $\log P[q|\sigma] = as_q(\sigma) + b$  for two constants  $a > 0, b$ . Thus in this case the optimal solution to the NSW problem is identical with the *maximum likelihood* order that is consistent with  $q$ . This in particular implies that given a prior uniform distribution on the  $n!$  rankings, any order  $\sigma$  maximizing (2) is also a maximizers of the posterior probability given  $q$ . So by analogy to problems in coding theory, see e.g. [7],  $\sigma$  is a maximum likelihood decoding of the original order  $\pi$ .

Note furthermore that one should not expect to be able to find the true order if  $q$  is noisy. Indeed for any pair of adjacent elements we are only given one noisy bit to determine which of the two is bigger.

## 1.2 Related Sorting Models and Results

It is natural to consider the problem of finding an a ranking  $\sigma$  that minimizes the score  $s_q(\sigma)$  without making any assumptions on the input  $q$ . This problem, called the *feedback arc set problem for tournaments* is known to be NP hard [1, 2]. However, it does admit PTAS [6] achieving a  $(1 + \epsilon)$  approximation for

$$-\frac{1}{2} \left[ s_q(\sigma) - \binom{n}{2} \right].$$

in time that is polynomial in  $n$  and doubly exponential in  $1/\epsilon$ . The results of [6] are the latest in a long line of work starting in the 1960's and including [1, 2]. See [6] for a detailed history of the feedback arc set problem.

A problem that is in a sense easier than NSW is the problem where repetitions are allowed in querying. In this case it is easy to observe that the original order may be recovered in  $O(n \log^2 n)$  queries with high probability. Indeed, one may perform any of the standard  $O(n \log n)$  sorting algorithms and repeat each query  $O(\log n)$  times in order to obtain the actual order between the queries elements with error probability  $n^{-2}$  (say). More sophisticated methods allow to show that in fact the true order may be found in query complexity  $O(n \log n)$  with high probability [4], see also [5].

## 1.3 Main Results

In our main results we show that the NSW problem is solvable in polynomial time with high probability and that any optimal order is close to the true order. More formally we show that

**Theorem 3.** *There exists a randomized algorithm that for any  $\gamma > 0$  and  $\beta > 0$  finds an optimal solution to the noisy sorting without resampling (NSWR) problem in time  $n^{O((\beta+1)\gamma^{-4})}$  except with probability  $n^{-\beta}$ .*

**Theorem 4.** *Consider the NSW problem and let  $\pi$  be the true order and  $\sigma$  be any optimal order then except with probability  $O(n^{-\beta})$  it holds that*

$$\sum_{i=1}^n |\sigma(i) - \pi(i)| = O(n), \tag{3}$$

$$\max_i |\sigma(i) - \pi(i)| = O(\log n). \tag{4}$$

Utilizing some of the techniques of [4] it is possible to obtain the results of Theorem 3 with low sampling complexity. More formally,

**Theorem 5.** *There is an implementation of a sorting algorithm with the same guarantees as in Theorem 3 and whose sampling complexity is  $C n \log n$  where  $C = C(\beta, \gamma)$ .*

It should be noted that the proofs can be modified to a more general case where the conditional probability from (1) is always bounded from below by  $p$  without necessarily being independent.

## 1.4 Techniques

In order to obtain a polynomial time algorithm for the NSW problem is important to identify that any optimal solution to the problem is close to the true one. Thus the main step of the analysis is the proof of Theorem 4.

To find efficient sorting we use an insertion algorithm. Given an optimal order on a subset of the items we show how to insert a new element. Since the optimal order both before and after the insertion of the element has to satisfy Theorem 4, it is also the case that no element moves more than  $O(\log n)$  after the insertion and re-sorting. Using this and a dynamic programming approach we derive an insertion algorithm in Section 2. The results of this section may be of independent interest in cases where it is known that a single element insertion into an optimal suborder cannot result in a new optimal order where some elements moved by much.

The main task is to prove Theorem 4 in Section 3. We first prove (3) by showing that for a large enough constant  $c$ , it is unlikely that any order  $\sigma$  whose total distance is more than  $cn$  will have  $s_q(\sigma) \geq s_q(\pi)$ , where  $\pi$  is the original order. We then establish (4) in subsection 3.2 using a bootstrap argument. The argument is based on the idea that if the discrepancy in the position of an element  $a$  in an optimal order compared to the true order is more than  $c \log n$  for a large constant  $c$ , then there must exist many elements that are “close” to  $a$  that have also moved by much. This then leads to a contradiction with (3).

The final analysis of the insertion algorithm and the proof of Theorem 3 are provided in Section 4. Section 5 shows how using a variant of the sorting algorithm it is possible to achieve polynomial running time in sampling complexity  $O(n \log n)$ .

## 1.5 Distances between rankings

Here we define a few measures of distance between rankings that will be used later. First, given two permutations  $\sigma$  and  $\tau$  we define the *dislocation distance* by

$$d(\sigma, \tau) = \sum_{i=1}^n |\sigma(i) - \tau(i)|.$$

Given a ranking  $\pi$  we define  $q_\pi \in \{\pm\}^{\binom{[n]}{2}}$  so that  $q_\pi(a_i, a_j) = +$  if  $\pi(i) > \pi(j)$  and  $q_\pi(a_i, a_j) = -$  otherwise. Note that using this notation  $q$  is obtained from  $q_\pi$  by flipping each entry independently with probability  $1 - p = 1/2 - \gamma$ . Given  $q, q' \in \{\pm\}^{\binom{[n]}{2}}$  we denote by

$$d(q, q') = \frac{1}{2} \sum_{i < j} |q(i, j) - q'(i, j)|$$

We will write  $d(\sigma)$  for  $d(\sigma, id)$  where  $id$  is the identity permutation and  $d(q)$  for  $d(q, q_{id})$ . Below we will often use the following well known claim [3].

**Claim 6.** *For any  $\tau$ ,*

$$\frac{1}{2}d(\tau) \leq d(q_\tau) \leq d(\tau).$$

## 2 Sorting a presorted list

In this section we prove that if a list is pre-sorted so that each element is at most  $k$  positions away from its location in the optimal ordering, then the optimal sorting can be found in time  $O(n^2 \cdot 2^{6k})$ .

**Lemma 7.** *Let  $a_1, a_2, \dots, a_n$  be  $n$  elements together with noisy queries  $q$ . Suppose that we are given that there is an optimal ordering  $a_{\sigma(1)}, a_{\sigma(2)}, \dots, a_{\sigma(n)}$ , such that  $|\sigma(i) - i| \leq k$  for all  $i$ . Then we can find such an optimal  $\sigma$  in time  $O(n^2 \cdot 2^{6k})$ .*

In the applications below  $k$  will be  $O(\log n)$ . Note that a brute force search over all possible  $\sigma$  would require time  $k^{\Theta(n)}$ . Instead we use dynamic programming to reduce the running time.

*Proof.* We use a dynamic programming technique to find an optimal sorting. In order to simplify notation we assume that the true ranking  $\pi$  is the identity ranking. In other words,  $a_1 < a_2 < \dots < a_n$ . Let  $i < j$  be any indices, then by the assumption, the elements in the optimally ordered interval

$$I = [a_{\sigma(i)}, a_{\sigma(i+1)}, \dots, a_{\sigma(j)}]$$

satisfy  $I^- \subset I \subset I^+$  where

$$I^+ = [a_{i-k}, a_{i-k+1}, \dots, a_{j+k}], \quad I^- = [a_{i+k}, a_{i+k+1}, \dots, a_{j-k}].$$

Hence selecting the set  $S_I = \{a_{\sigma(i)}, a_{\sigma(i+1)}, \dots, a_{\sigma(j)}\}$  involves choosing a set of size  $j - i + 1$  that contains the elements of  $I^-$  and is contained in  $I^+$ . This involves selecting  $2k$  elements from the list (or from a subset of the list)

$$\{a_{i-k}, a_{i-k+1}, \dots, a_{i+k-1}, a_{j-k+1}, a_{j-k+2}, \dots, a_{j-k}\}$$

which has  $4k$  elements. Thus the number of such  $S_I$ 's is bounded by  $2^{4k}$ .

We may assume without loss of generality that  $n$  is an exact power of 2. Denote by  $I_0$  the interval containing all the elements. Denote by  $I_1$  the left half of  $I_0$  and by  $I_2$  its right half. Denote by  $I_3$  the left half of  $I_1$  and so on. In total, we will have  $n - 1$  intervals of lengths  $2, 4, 8, \dots$

For each  $I_t = [a_i, \dots, a_j]$  let  $S_t$  denote the possible ( $< 2^{4k}$ ) sets of the elements  $I'_t = [a_{\sigma(i)}, \dots, a_{\sigma(j)}]$ . We use dynamic programming to store an optimal ordering of each such  $I'_t \in S_t$ . The total number of  $I'_t$ 's we will have to consider is bounded by  $n \cdot 2^{4k}$ . We proceed from  $t = n - 1$  down to  $t = 0$  producing and storing an optimal sort for each possible  $I'_t$ . For  $t = n - 1, n - 2, \dots, n/2$  the length of each  $I'_t$  is 2, and the optimal sort can be found in  $O(1)$  steps.

Now let  $t < n/2$ . We are trying to find an optimal sort of a given  $I'_t = [i, i + 2s - 1]$ . We do this by dividing the optimal sort into two halves  $I_l$  and  $I_r$  and trying to sort them separately. We know that  $I_l$  must contain all the elements in  $I'_t$  that come from the interval  $[a_1, \dots, a_{i+s-1-k}]$  and must be contained in the interval  $[a_1, \dots, a_{i+s-1+k}]$ . Thus there are at most  $2^{2k}$  choices for the elements of  $I_l$ , and the choice of  $I_l$  determines  $I_r$  uniquely. For each such choice we look up an optimum solution for  $I_l$  and for  $I_r$  in the dynamic programming table. Among all possible choices of  $I_l$  we pick the best one. This is done by recomputing the score  $s_q$  for the joined interval, and takes at most  $|I'_t|^2$  time. Thus the total cost will be

$$\sum_{i=1}^{\log n} \# \text{intervals of length } 2^i \cdot \# \text{checks} \cdot \text{cost of check} = \sum_{i=1}^{\log n} O\left(\frac{n \cdot 2^{4k}}{2^i} \cdot 2^{2k} \cdot 2^{2i}\right) = O(n^2 \cdot 2^{6k}).$$

□

### 3 The Discrepancy between the true order and Optima

The goal of this section is to establish that with high probability any optimum solution will not be far from the original solution. We first establish that the orders are close on average, and then that they are pointwise close to each other.

#### 3.1 Average proximity

We prove that with high probability, the total difference between the original and any optimal ordering is linear in the length of the interval.

We begin by bounding the probability that a specific permutation  $\sigma$  will beat the original ordering.

**Lemma 8.** *Suppose that the original ordering is  $a_1 < a_2 \dots < a_n$ . Let  $\sigma$  be another permutation. Then the probability that  $\sigma$  beats the identity permutation is bounded from above by*

$$P[\text{Bin}(d(q_\sigma), 1/2 + \gamma) \leq d(q_\sigma)/2] \leq \exp(-2d(q_\sigma)\gamma^2)$$

*Proof.* In order for  $\sigma$  to beat the identity, it needs to beat it in at least half of the  $d(q_\sigma)$  pairwise relation where they differ. This proves that the probability that it beats the identity is exactly  $P[\text{Bin}(d(q_\sigma), 1/2 + \gamma) \leq d(q_\sigma)/2]$ . The last inequality follows by a Chernoff bound.  $\square$

**Lemma 9.** *The number of permutations  $\tau$  on  $[n]$  satisfying  $d(\tau) \leq cn$  is at most*

$$2^n 2^{(1+c)n H(1/(1+c))}.$$

Here  $H(x)$  is the binary entropy of  $x$  defined by

$$H(x) = -x \log_2 x - (1-x) \log_2 (1-x) < -2x \log_2 x,$$

for small  $x$ .

*Proof.* Note that each  $\tau$  can be uniquely specified by the values of  $s(i) = \tau(i) - i$ , that we are given that  $\sum |s(i)|$  is exactly  $d(\tau) \leq cn$ . Thus there is an injection of  $\tau$ 's with  $d(\tau) = m$  into sequences of  $n$  numbers which in absolute values add up to  $m$ . It thus suffices to bound the number of such sequences. The number of unsigned sequences equals the number of ways of placing  $m$  balls in  $n$  bins, which is equal to  $\binom{n+m-1}{n-1}$ . Signs multiply the possibilities by at most  $2^n$ . Hence the total number of  $\tau$ 's with  $d(\tau) = m$  is bounded by  $2^n \cdot \binom{n+m-1}{n-1}$ . Summing up over the possible values of  $m$  we obtain

$$\sum_{m=0}^{cn} 2^n \cdot \binom{n+m-1}{n-1} < 2^n \cdot \binom{n+cn}{n} \leq 2^n 2^{(n+cn) H(n/(n+cn))}. \quad (5)$$

$\square$

**Lemma 10.** *Suppose that the true ordering is  $a_1 < \dots < a_n$  and  $n$  is large enough. Then if  $c \geq 1$  and*

$$\gamma^2 c > 1 + (1+c)H(1/(1+c)),$$

*the probability that any ranking  $\sigma$  is optimal and  $d(\sigma) > cn$  is at most  $\exp(-cn\gamma^2/10)$  for sufficiently large  $n$ . In particular, as  $\gamma \rightarrow 0$ , it suffices to take*

$$c = O(-\gamma^{-2} \log \gamma) = \tilde{O}(\gamma^{-2}).$$

*Proof.* Let  $\sigma$  be an ordering with  $d(\sigma) > cn$ . Then by Claim 6 we have  $d(q_\sigma) > cn/2$ . Therefore the probability that such an ordering will beat the identity is bounded by  $\exp(-cn\gamma^2)$  by Lemma 8. We now use union bound and Lemma 9 to obtain the desired result.  $\square$

### 3.2 Pointwise proximity

In the previous section we have seen that it is unlikely that the *average* element in the optimal order is more than a constant number of positions away from its original location. Our next goal is to show that the *maximum* dislocation of an element is bounded by  $O(\log n)$ . As a first step, we show that one “big” dislocation is likely to entail many “big” dislocations.

**Lemma 11.** *Suppose that the true ordering of  $a_1, \dots, a_n$  is given by the identity ranking, i.e.,  $a_1 < a_2 < \dots < a_n$ . Let  $1 \leq i < j \leq n$  be two indices and  $m = j - i$ . Let  $A_{ij}$  be the event that there is an optimum ordering  $\sigma$  such that  $\sigma(i) = j$  and*

$$(\sigma[1, i - \ell - 1] \cup \sigma[j + \ell + 1, n]) \cap [i, j - 1] \leq \ell,$$

*i.e., at most  $\ell$  elements are mapped to the interval  $[i, j - 1]$  from outside the interval  $[i - \ell, j + \ell]$  by  $\sigma$ , where  $\ell = \lfloor \frac{1}{6}\gamma m \rfloor$ . Then*

$$P(A_{ij}) < p_1^m,$$

*where  $p_1 = \exp(-\gamma^2/16) < 1$ .*

*Proof.* The assumption that  $\sigma$  is optimal implies in particular that moving the  $i$ -th element from the  $j$ -th position where it is mapped by  $\sigma$  back to the  $i$ -th position does not improve the solution. The event  $A_{ij}$  implies that among the elements  $a_k$  for  $k \in [i - \ell, j + \ell]$  at least  $m/2 - \ell$  satisfy  $q(k, i) = -$ . This means that at least

$$\frac{m}{2} - 2\ell - 1 > \frac{m}{2} - \frac{\gamma}{2}m + \frac{\ell}{2} > \left(\frac{1}{2} - \frac{\gamma}{2}\right)(m + \ell)$$

of the elements  $a_k$  for  $k \in [i + 1, j + \ell]$  must satisfy  $q(k, i) = -$ . The probability of this occurring is less than

$$\exp\left(\frac{-\frac{m+\ell}{2}(\gamma/2)^2}{2}\right) = p_1^{m+\ell}$$

using Chernoff bounds. □

As a corollary to Lemma 11 we obtain the following using a simple union-bound. For the rest of the proof all the log's are base 2.

**Corollary 12.** *Let*

$$m_1 = (-\log \varepsilon + 2 \log n / \log(1/p_1)) = O((-\log \varepsilon + \log n)/\gamma^2),$$

*then  $A_{ij}$  does not occur for any  $i, j$  with  $|i - j| \geq m_1$  with probability  $> 1 - \varepsilon$ .*

Next, we formulate a corollary to Lemma 10.

**Corollary 13.** *Suppose that  $a_1 < a_2 < \dots < a_n$  is the true ordering. Set  $m_2 = 2m_1$ . For each interval  $I = [a_i, \dots, a_j]$  with at least  $m_2$  elements consider all the sets  $S_I$  which contain the elements from*

$$I^- = [a_{i+m_2}, \dots, a_{j-m_2}],$$

*and are contained in the interval*

$$I^+ = [a_{i-m_2}, \dots, a_{j+m_2}].$$



Then with probability  $> 1 - \varepsilon$  all such sets  $S_I$  do not have an optimal ordering that has a total deviation from the true of more than  $c_2 |i - j|$ , with

$$c_2 = \frac{70}{\gamma^2} = O(\gamma^{-2}),$$

a constant.

*Proof.* There are at most  $n^2 \cdot 2^{4m_2}$  such intervals. The probability of each interval not satisfying the conclusion is bounded by Lemma 10 with

$$e^{-c_2 m_2 \gamma^2 / 10} = e^{-7m_2} < 2^{-7m_2} = 2^{-m_2} \cdot 2^{-2m_2} \cdot 2^{-4m_2} < \varepsilon \cdot n^{-2} \cdot 2^{-4m_2}.$$

The last inequality holds because  $m_2 > \max(\log n, -\log \varepsilon)$ . By taking a union bound over all the sets we obtain the statement of the corollary.  $\square$

We are now ready to prove the main result on the pointwise distance between an optimal ordering and the original.

**Lemma 14.** *Assuming that the events from Corollaries 12 and 13 hold, it follows that for each optimal ordering  $\sigma$  and for each  $i$ ,  $|i - \sigma(i)| < c_3 \log n$ , where*

$$c_3 = 500 \gamma^{-2} \cdot \frac{m_2}{\log n} = O(\gamma^{-4}(-\log \varepsilon / \log n + 1))$$

is a constant. In particular, this conclusion holds with probability  $> 1 - 2\varepsilon$ .

*Proof.* Assume that the events from both corollaries hold, and let  $\sigma$  be an optimal ordering. We say that a position  $i$  is *good* if there is no index  $j$  such that  $\sigma(j)$  is on the other side of  $i$  from  $j$  and  $|\sigma(j) - j| \geq m_2$ . In other words,  $i$  is good if there is no "long" jump over  $i$  in  $\sigma$ . In the case when  $i = j$  or  $i = \sigma(j)$  for a long jump, it is not considered good. An index that is not good is bad. An interval  $I$  is bad if all of its indices are bad. Our goal is to show that there are no bad intervals of length  $\geq c_3 \log n$ . This would prove the lemma, since if there is an  $i$  with  $|i - \sigma(i)| > c_3 \log n$  then there is a bad interval of length at least  $c_3 \log n$ .

Assume, for contradiction, that  $I = [i, \dots, i + t - 1]$  is a bad interval of length  $t \geq c_3 \log n$ , such that  $i - 1$  and  $i + t$  are both good (or lie beyond the endpoints of  $[1, \dots, n]$ ). Denote by  $S$  the set of elements that is mapped to  $I$  by  $\sigma$ . Denote the indices in  $S$  in their original order by  $i_1 < i_2 < \dots < i_t$ , i.e., we have:  $\{\sigma(i_1), \dots, \sigma(i_t)\} = I$ .

By the goodness of the endpoints of  $I$  we have

$$[i + m_2, i + t - 1 - m_2] \subset \{i_1, \dots, i_t\} \subset [i - m_2, i + t - 1 + m_2].$$

Denote the permutation induced by  $\sigma$  on  $S$  by  $\sigma'$  so  $\sigma(i_j) < \sigma(i_{j'})$  is equivalent to  $\sigma'(j) < \sigma'(j')$ . The permutation  $\sigma'$  is optimal, for otherwise it would have been possible to improve  $\sigma$  by improving  $\sigma'$ .

By Corollary 13 and Claim 6, we have

$$d(q_{\sigma'}) \leq d(\sigma') \leq c_2 t.$$

In how many switches can the elements of  $S$  participate under  $\sigma$ ? They participate in switches with other elements of  $S$  to a total of  $d(q_{\sigma'})$ . In addition, they participate in switches with elements that are not in  $S$ . These elements must originate at the margins of the interval  $i$ : either in the interval  $[i - m_2, i + m_2]$  or the



interval  $[i + t - 1 - m_2, i + t - 1 + m_2]$ . Thus, each contributes at most  $2m_2$  switches with elements of  $S$ . There are at most  $2m_2$  such elements. Hence the total number of switches between elements in  $S$  and in  $\bar{S}$  is at most  $4m_2^2$ . Hence

$$\sum_{i \in S} |\sigma(i) - i| \leq \sum_{i \in S} \#\{\text{switches } i \text{ participates in}\} \leq 4m_2^2 + 2d(q_{\sigma'}) \leq 4m_2^2 + 2c_2t. \quad (6)$$

We assumed that the entire interval  $I$  is bad, hence for every position  $i$  there is an index  $j_i$  such that  $|\sigma(j_i) - j_i| \geq m_2$  and such that  $i$  is in the interval  $J_i = [j_i, \sigma(j_i)]$  (or the interval  $[\sigma(j_i), j_i]$ , depending on the order). Consider all such  $J_i$ 's. By a Vitali covering lemma argument we can choose a disjoint collection of them whose total length is at least  $|I|/3$ . The argument proceeds as follows: Order the intervals in a decreasing length order (break ties arbitrarily). Go through the list and add a  $J_i$  to our collection if it is disjoint from all the currently selected intervals. We obtain a collection  $J_1, \dots, J_k$  of disjoint intervals of the form  $[j_i, \sigma(j_i)]$ . Denote the length of the  $i$ -th interval by  $t_i = |j_i - \sigma(j_i)|$ . Let  $J'_i$  be the "tripling" of the interval  $J_i$ :  $J'_i = [j_i - t_i, \sigma(j_i) + t_i]$ . We claim that the  $J'_i$ -s cover the entire interval  $I$ . Let  $m$  be a position on the interval  $I$ . Then there is an interval of the form  $[j, \sigma(j)]$  (or  $[\sigma(j), j]$ ) that covers  $m$ . Choose the longest such interval  $J' = [j, \sigma(j)]$ . If  $J'$  has been selected to our collection then we are done. If not, it means that  $J'$  intersects a longer interval  $J_i$  that has been selected. This means that  $J'$  is covered by the tripled interval  $J'_i$ . In particular,  $m$  is covered by  $J'_i$ . We conclude that

$$t = \text{length}(I) \leq \sum_{i=1}^k \text{length}(J'_i) = 3 \sum_{i=1}^k t_i.$$

Thus  $\sum_{i=1}^k t_i \geq t/3$ . This concludes the covering argument.

We now apply Corollary 12 to the intervals  $J_i$ . We conclude that on an interval  $J_i$  the contribution of the elements of  $S$  that are mapped to  $J_i$  to the sum of deviations under  $\sigma$  is at least  $\ell_i^2$  where  $\ell_i = \frac{1}{6}\gamma t_i$ . Thus

$$\begin{aligned} \sum_{i \in S} |\sigma(i) - i| &\geq \sum_{j=1}^k \ell_j^2 = \frac{1}{36}\gamma^2 \cdot \sum_{j=1}^k t_j^2 \geq \frac{1}{36}\gamma^2 \cdot m_2 \cdot \sum_{j=1}^k t_j \\ &\geq \frac{1}{36}\gamma^2 \cdot m_2 \cdot t/3 \geq m_2 \cdot \frac{1}{125}\gamma^2 \cdot c_3 \log n + \frac{1}{800}\gamma^2 \cdot m_2 t \\ &> m_2 \cdot (4m_2) + 2c_2t = 4m_2^2 + 2c_2t, \end{aligned}$$

for sufficiently large  $n$ . The result contradicts (6) above. Hence there are no bad intervals of length  $\geq c_3 \log n$ , which completes the proof.  $\square$

## 4 The algorithm

We are now ready to give an algorithm for computing the optimal ordering with high probability in polynomial time. Note that Lemma 14 holds for any interval of length  $\leq n$  (not just length exactly  $n$ ). Set  $\varepsilon = n^{-\beta-1}/4$ . Given an input, let  $S \subset \{a_1, \dots, a_n\}$  be a random set of size  $k$ . The probability that there is an optimal ordering  $\sigma$  of  $S$  and an index  $i$  such that  $|i - \sigma(i)| \geq c_3 \log n$ , where

$$c_3 = O(\gamma^{-4}(-\log \varepsilon / \log n + 1)) = O(\gamma^{-4}(\beta + 1)),$$

is bounded by  $2\varepsilon$  by Lemma 14. Let

$$S_1 \subset S_2 \subset \dots \subset S_n$$

be a randomly selected chain of sets such that  $|S_k| = k$ . Then the probability that an element of an optimal order of any of the  $S_k$ 's deviates from its original location by more than  $c_3 \log n$  is bounded by  $2n\varepsilon = n^{-\beta}/2$ . We obtain:

**Lemma 15.** *Let  $S_1 \subset \dots \subset S_n$  be a chain of randomly chosen subsets with  $|S_k| = k$ . Denote by  $\sigma_k$  an optimal ordering on  $S_k$ . Then with probability  $\geq 1 - n^{-\beta}/2$ , for each  $\sigma_k$  and for each  $i$ ,  $|i - \sigma_k(i)| < c_3 \log n$ , where  $c_3 = O(\gamma^{-4}(\beta + 1))$  is a constant.*

We are now ready to prove the main result.

**Theorem 16.** *There is an algorithm that runs in time  $n^{c_4}$  where*

$$c_4 = O(\gamma^{-4}(\beta + 1))$$

*is a constant that outputs an optimal ordering with probability  $\geq 1 - n^{-\beta}$ .*

*Proof.* First, we choose a random chain of sets  $S_1 \subset \dots \subset S_n$  such that  $|S_k| = k$ . Then by Lemma 15, with probability  $1 - n^{-\beta}/2$ , for each optimal order  $\sigma_k$  of  $S_k$  and for each  $i$ ,  $|i - \sigma_k(i)| < c_3 \log n$ . We will find the orders  $\sigma_k$  iteratively until we reach  $\sigma_n$  which will be an optimal order for our problem. Denote  $\{a_k\} = S_k - S_{k-1}$ . Suppose that we have computed  $\sigma_{k-1}$  and we would like to compute  $\sigma_k$ . We first insert  $a_k$  into a location that is close to its original location as follows. Break  $S_k$  into blocks  $B_1, B_2, \dots, B_s$  of length  $c_3 \log n$ . We claim that with probability  $> n^{-\beta-1}/2$  we can pinpoint the block  $a_k$  belongs to within an error of  $\pm 2$ , thus locating  $a_k$  within  $3c_3 \log n$  of its original location.

Suppose that  $a_k$  should belong to block  $B_i$ . Then by our assumption on  $\sigma_{k-1}$ ,  $a_k$  is bigger than any element in  $B_1, \dots, B_{i-2}$  and smaller than any element in  $B_{i+2}, \dots, B_s$ . By comparing  $a_k$  to each element in the block and taking majority, we see that the probability of having an incorrect comparison result with a block  $B_j$  is bounded by  $n^{-\beta-2}/2$ . Hence the probability that  $a_k$  will not be placed correctly up to an error of two blocks is bounded by  $n^{-\beta-1}/2$  using union bound.

Hence after inserting  $a_k$  we obtain an ordering of  $S_k$  in which each element is at most  $3c_3 \log n$  positions away from its original location. Hence each element is at most  $4c_3 \log n$  positions away from its optimal location in  $\sigma_k$ . Thus, by Lemma 7 we can obtain  $\sigma_k$  in time  $O(n^{24c_3+2})$ . The process is then repeated.

The probability of each stage failing is bounded by  $n^{-\beta-1}/2$ . Hence the probability of the algorithm failing assuming the chain  $S_1 \subset \dots \subset S_n$  satisfies Lemma 15 is bounded by  $n^{-\beta}/2$ . Thus the algorithm runs in time  $O(n^{24c_3+3})$  and has a failure probability of at most  $n^{-\beta}/2 + n^{-\beta}/2 = n^{-\beta}$ .  $\square$

## 5 Query Complexity

Here we briefly sketch the proof of Theorem 5. Recall that the theorem states that although the running time of the algorithm is a polynomial of  $n$  whose degree depends on  $p$ , the query complexity of a variant of the algorithm is  $O(n \log n)$ . Note that there are two types of queries. The first type is comparing elements in the dynamic programming, while the second is when inserting new elements.

**Lemma 17.** *For all  $\beta > 0, \gamma < 1/2$  there exists  $c(\beta, \gamma) < \infty$  such that the total number of comparisons performed in the dynamic programming stage is  $O(n \log n)$  of the algorithm is at most  $c n \log n$  except with probability  $O(n^{-\beta})$ .*

*Proof.* Recall that in the dynamic programming stage, each element is compared with elements that are at current distance at most  $c_0 \log n$  from it where  $c_0 = c_0(\beta, \gamma)$ .

Consider a random insertion order of the elements  $a_1, \dots, a_n$ . Let  $S_{n/2}$  denote the set of elements inserted up to the  $n/2$  insertion. Then by standard concentration results it follows that there exists  $c_1(c_0, \beta)$  such that for all  $1 \leq i \leq n - c_1 \log n$  it holds that

$$|[a_i, a_i + c_1 \log n] \cap S_{n/2}| \geq c_0 \log n, \quad (7)$$

and for all  $c_1 \log n \leq i \leq n$  it holds that

$$|[a_i - c_1 \log n, a_i] \cap S_{n/2}| \geq c_0 \log n \quad (8)$$

except with probability at most  $n^{-\beta-1}$ . Note that when (7) and (8) both hold the number of different queries used in the dynamic programming while inserting the elements in  $\{a_1, \dots, a_n\} \setminus S_{n/2}$  is at most  $2c_1 n \log n$ .

Repeating the argument above for the insertions performed from  $S_{n/4}$  to  $S_{n/2}$ , from  $S_{n/8}$  to  $S_{n/4}$  etc. we obtain that the total number of queries used is bounded by:

$$2c_1 \log n (n + n/2 + \dots + 1) \leq 4c_1 n \log n,$$

except with probability  $2n^{-\beta}$ . This concludes the proof.  $\square$

Next we show that there is implementation of insertion that requires only  $O(\log n)$  comparisons per insertion.

**Lemma 18.** *For all  $\beta > 0$  and  $\gamma < 1/2$  there exists a  $C(\beta, \gamma) = O(\gamma^{-2}(\beta + 1))$  and  $c(\beta, \gamma) = O(\gamma^{-4}(\beta + 1))$  such that except with probability  $O(n^{-\beta})$  it is possible to perform the insertion in the proof of Theorem 16 so that each element is inserted using at most  $C \log n$  comparisons,  $O(\log n)$  time and the element is placed a distance of at most  $c \log n$  from its optimal location.*

*Proof.* Below we assume (as in the proof of Theorem 16) that there exists  $c_1(\beta, \gamma) = O(\gamma^{-4}(\beta + 1))$  such that at all stages of the insertion and for each item, the distance between the location of the item in the original order and the optimal order is at most  $c_1 \log n$ . This will result in an error with probability at most  $n^{-\beta}/2$ . Let  $k = k(\gamma) = O(\gamma^{-2})$  be a constant such that

$$P[\text{Bin}(k, 1/2 + \gamma) > k/2] > 1 - 10^{-3}.$$

Let  $c_2 = O(\beta + 1)$  be chosen so that

$$P[\text{Bin}(c_2 \log n, 0.99) < \frac{c_2}{2} \log n + 2 \log_2 n] < n^{-\beta-1}, \quad (9)$$

Let  $c_3 = kc_2 + 4c_1$ .

We now describe an insertion step. Let  $S$  denote a currently optimally sorted set. We will partition  $S$  into consecutive intervals of length between  $c_3 \log n$  and  $2c_3 \log n$  denoted  $I_1, \dots, I_t$ . We will use the notation  $I'_i$  for the sub-interval of  $I_i = [s, t]$  defined by  $I'_i = [s + 2c_1 \log n, t - 2c_1 \log n]$ . We say that a newly inserted element  $a_j$  belongs to one of the interval  $I_i$  if one of the two closest elements to it in the original order belongs to  $I_i$ . Note that  $a_j$  can belong to at most two intervals. An element in  $S$  belongs to  $I_i$  iff it is one of the elements in  $I_i$ . Note furthermore that if  $a_j$  belongs to the interval  $I_i$  then its optimal insertion location is determined up to  $2(kc_2 + 6c_1) \log n$ . Similarly, if we know it belongs to one of two

intervals then its optimal insertion location is determined up to  $4(kc_2 + 6c_1) \log n$ , therefore we can take  $c = 4(kc_2 + 6c_1) = O(\gamma^{-4}(\beta + 1))$ .

Note that by the choice of  $c_1$  we may assume that all elements belonging to  $I_i$  are smaller than all elements of  $I'_j$  if  $i < j$  in the true order. Similarly, all elements belonging to  $I_j$  are larger than all elements of  $I'_i$  if  $j > i$ . We define formally the interval  $I_0 = I'_0$  to be an interval of elements that are smaller than all the items and the interval  $I_{t+1} = I'_{t+1}$  to be an interval of elements that is bigger than all items.

We construct a binary search tree on the set  $[1, t]$  labeled by sub-intervals of  $[1, t]$  such that the root is labeled by  $[1, t]$  and if a node is labeled by an interval  $[s_1, s_2]$  with  $s_2 - s_1 > 1$  then its two children are labeled by  $[s_1, s']$  and  $[s', s_2]$ , where  $s'$  is chosen so that the length of the two intervals is the same up to  $\pm 1$ . Note that the two sub-interval overlap at  $s'$ . This branching process terminates at intervals of the form  $[s, s + 1]$ . Each such node will have a path of descendants of length  $c_2 \log n$  all labeled by  $[s, s + 1]$ .

We will use a variant of binary insertion closely related to the algorithm described in Section 3 of [4]. The algorithm will run for  $c_2 \log n$  steps starting at the root of the tree. At each step the algorithm will proceed from a node of the tree to either one of the two children of the node or to the parent of that node.

Suppose that the algorithm is at the node labeled by  $[s_1, s_2]$  and  $s_2 - s_1 > 1$ . The algorithm will first take  $k$  elements from  $I'_{s_1-1}$  that have not been explored before and will check that the current item is greater than the majority of them. Similarly, it will make a comparison with  $k$  elements from  $I'_{s_2+1}$ . If either test fails it would backtrack to the parent of the current node. Note that if the test fails then it is the case that the element does not belong to  $[s_1, s_2]$  except with probability  $10^{-2}$ .

Otherwise, let  $[s_1, s']$  and  $[s', s_2]$  denote the two children of  $[s_1, s_2]$ . The algorithm will now perform a majority test against  $k$  elements from  $I_{s'}$  according to which it would choose one of the two sub-interval  $[s_1, s']$  or  $[s', s_2]$ . Note again that a correct sub-interval is chosen except with probability at most  $10^{-2}$  (note that in this case there may be two “correct” intervals).

In the case where  $s_2 = s_1 + 1$  we perform only the first test. If it fails we move to the parent of the node. If it succeeds, we move to the single child. Again, note that we will move toward the leaf if the interval is correct with probability at least 0.99. Similarly, we will move away from the leaf if the interval is incorrect with probability at least 0.99.

Overall, the analysis shows that at each step we move toward a leaf including the correct interval with probability at least 0.99. From (9) it follows that with probability at least  $1 - n^{-\beta-1}$  after  $c_2 \log n$  steps the label of the current node will be  $[s, s + 1]$  where the inserted element belongs to either  $I_s$  or  $I_{s+1}$ . Thus the total number of queries is bounded by  $3kc_2 \log n$  and we can take  $C = 3kc_2 = O(\gamma^{-2}(\beta + 1))$ . This concluded the proof.  $\square$

## References

- [1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. In *Proceedings of 37rd STOC*, 2005.
- [2] N. Alon. Ranking tournaments. *Siam Journal on Discrete Mathematics*, 20(1):137–142, 2006.
- [3] P. Diaconis and R. L. Graham. Spearman’s footrule as a measure of disarray. *J. Roy. Statist. Soc. Ser. B*, 39(2):262–268, 1977.
- [4] U. Feige, D. Peleg, P. Raghavan, and E. Upfal. Computing with unreliable information. In *Proceedings 22nd STOC*, 1990.
- [5] D. Karp and B. Kleinberg. Noisy binary search and its applications. In *Proceedings of SODA*, pages 891–890, 2007.
- [6] C. Kenyon-Mathieu and W. Schudy. How to rank with few errors. In *STOC*, pages 95–103, 2007.
- [7] S. Romann. *Introduction to Coding and Information Theory*. Springer, 1997.